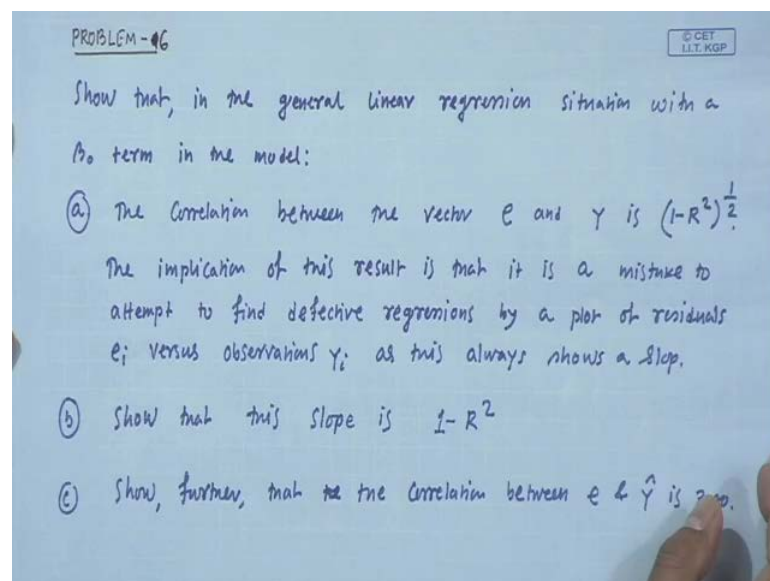


Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 38
Tutorial – III

This is my third tutorial, and today we will be solving you know some randomly selected problems.

(Refer Slide Time: 00:30)



Let me start with this problem, it is problem 6, because I have solved the 5 problems before. So that in the general linear regression situation with the beta naught term in the model, the correlation between the vector e and Y is 1 minus R square square root of this. This is sort of you know theoretical problem, but this has some implication, we will come to that point later on.

So, the first part says that you know you find the correlation between the vector e , so this e is the residual vector and Y is observed response. And we have to find the correlation between e and Y , and we have to prove that that is square root of 1 minus R square, where R square is the coefficient of multiple determination. So, we know that this R square is equal to $S S$ regression by $S S$ total, so this R square parameter it sort of measure the proportion of variability in Y , that is explained by the model.

(Refer Slide Time: 02:46)

© CET
I.I.T. KGP

$$\text{Cov}(e, Y) = \frac{\sum (e_i - \bar{e})(Y_i - \bar{Y})}{\sqrt{\sum (e_i - \bar{e})^2 \sum (Y_i - \bar{Y})^2}} = (1 - R^2)^{\frac{1}{2}} \quad R^2 = \frac{SS_{\text{Reg}}}{SS_T}$$

$$\sum_{i=1}^n (e_i - \bar{e})(Y_i - \bar{Y}) = \sum_{i=1}^n e_i (Y_i - \bar{Y}) \quad \left. \begin{array}{l} \bar{e} = 0 \text{ if } \beta_0 \text{ is in the} \\ \text{model.} \end{array} \right\}$$

$$= \sum_{i=1}^n e_i Y_i = \underline{e'Y = e'e.}$$

$$Y = X\beta + \epsilon$$

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1} X'Y = HY, \quad \hat{Y} = HY$$

where $H = X(X'X)^{-1}X'$

So, let me solve the first part a the correlation between e and Y, so the correlation between e and Y that is nothing but summation e i minus e bar into Y i minus Y bar by summation e i minus e bar square into Y i minus Y bar square and the square root of whole thing. So, we have to prove that this is equal to 1 minus R square, where R square is S S regression by S S total not, so easy.

So, we can start with this numerator, e i minus e bar Y i minus Y bar i is from 1 to n, so can I write this one as just e i into Y i minus Y bar, because of the fact that e bar is equal to 0. If beta naught is in the model, because if beta naught is in the model then while differentiating the least square function s with respect to beta naught, we get summation e i equal to 0, and from there it is e bar is equal to 0.

So, you can write, this is equal to e i into Y i minus Y bar i is from 1 to n, and this can be written again this can be written as e i Y i, i is from 1 to n, because you know this summation e i is equal to 0. So, in matrix notation this can be, in vector notation this can be written as e prime Y or Y prime e, anyway what is e prime Y can I write this as e prime e, it is not trivial, so we need to check this part, whether this is equal to e prime e.

So, let me start with e prime e, and then I will prove that e prime e is equal to e prime y, so before that let me recall the multiple linear regression model Y equal to X beta plus epsilon. And here, we know that the parameter beta is estimated by beta hat, which is equal to X prime X inverse X prime Y, and then the fitted model is Y hat is equal to X

beta hat, this is the fitted model. And we know that beta hat is equal to $X'X^{-1}X'Y$, so I can write this one as X and then I plug the value of beta hat here, so $X'X^{-1}X'Y$.

And then I call this matrix, I call this H we this is called hat matrix, you know we know about this H matrix, so this is called hat matrix, because of the fact that, so finally what we got is that we, so \hat{Y} is equal to HY , because this hat matrix comes from Y to \hat{Y} , that is why it is called hat matrix. Anyway, so where H is equal to $X'X^{-1}X$, so \hat{Y} is equal to HY , now what is e ?

(Refer Slide Time: 08:56)

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

$$e'e = Y'(I - H)'(I - H)Y = Y'e$$

$$e'e = e'Y$$

$$H^2 = H$$

Then e is equal to the observed value minus the estimated value, so just now we have proved that \hat{Y} is equal to HY , so this one is equal to I minus H Y , so I is the identity matrix. So, now what we want is that, I want to prove this $e'e$ is equal to $e'Y$ that is what I want, now I have a formula for e in terms of hat matrix. So, let me start now $e'e$ I can write $e'e$ is now, in terms of hat matrix that is $Y' I$ minus $H' I$ minus H Y .

And it is known that you know this hat matrix is idempotent matrix; that means, H^2 is equal to H , and then $I - H$ is also idempotent, so what I can write is that this is Y' , this can be replaced by only $I - H$ Y . Now, $I - H$ Y is equal to e , so this is equal to $Y'e$, so we have proved that $e'e$ equal to $Y'e$.

(Refer Slide Time: 11:02)

$$\text{Cor}(e, Y) = \frac{\sum (e_i - \bar{e})(Y_i - \bar{Y})}{\sqrt{\sum (e_i - \bar{e})^2 \sum (Y_i - \bar{Y})^2}} = (1 - R^2)^{\frac{1}{2}} \quad R^2 = \frac{SS_{\text{Reg}}}{SS_T}$$

$$\sum_{i=1}^n (e_i - \bar{e})(Y_i - \bar{Y}) = \sum_{i=1}^n e_i (Y_i - \bar{Y}) \quad \left. \begin{array}{l} \bar{e} = 0 \text{ if } \beta_0 \text{ is in the} \\ \text{model.} \end{array} \right\}$$

$$= \sum_{i=1}^n e_i Y_i = \underline{e'Y} = e'e = \sum e_i^2 = SS_{\text{Res}}$$

$$Y = X\beta + \epsilon$$

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1} X'Y = HY, \quad \hat{Y} = HY$$

where $H = X(X'X)^{-1} X'$

So, what we have proved is that the numerator is equal to $e'Y$, which is equal to $e'e$ is nothing but summation e_i^2 which is nothing but SS_{Residual} , so we have proved that the numerator is SS_{Residual} . And what is this, this is nothing but $e'Y$ and this one is nothing but $Y'Y$ I mean this is nothing but SS_{Total} , in fact let me write down once more here.

(Refer Slide Time: 11:56)

$$e = Y - \hat{Y} = Y - HY = (I - H)Y \quad H^2 = H$$

$$\boxed{e'e = e'Y} \quad e'e = Y'(I - H)'(I - H)Y$$

$$= Y'(I - H)Y = Y'e$$

$$\text{Cor}(e, Y) = \frac{\sum (e_i - \bar{e})(Y_i - \bar{Y})}{\sqrt{\sum (e_i - \bar{e})^2 \sum (Y_i - \bar{Y})^2}} = \frac{e'e}{\sqrt{(e'e) SS_T}}$$

$$= \sqrt{\frac{e'e}{SS_T}} = \sqrt{\frac{SS_{\text{Res}}}{SS_T}} = \sqrt{\frac{SS_T - SS_{\text{Reg}}}{SS_T}}$$

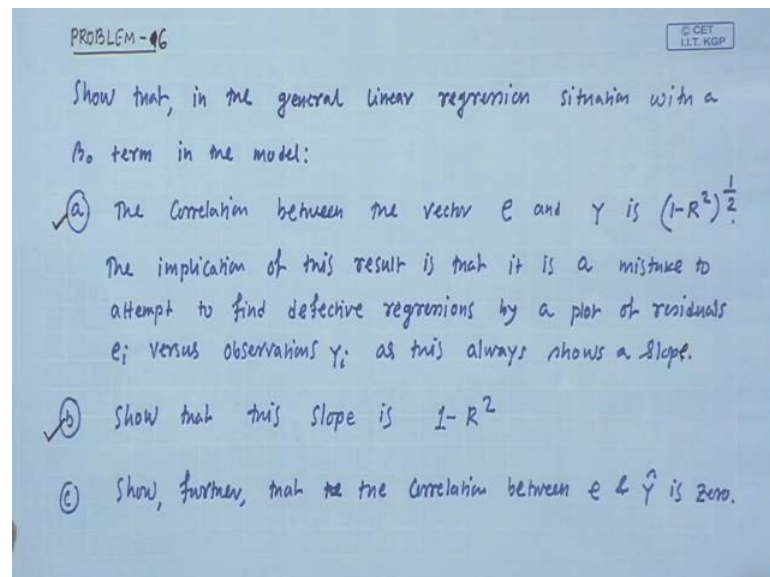
$$= \sqrt{1 - \frac{SS_{\text{Reg}}}{SS_T}} = (1 - R^2)^{\frac{1}{2}}$$

So, you want to find the correlation between e and Y , which is equal to summation e_i minus e bar Y_i minus Y bar by square root of summation e_i minus e bar square Y_i

minus \bar{Y} square. And we proved that this one is equal to $e' e$ the numerator is $e' e$ and then since \bar{e} is equal to 0, this one is nothing but $\sum e_i^2$, so this is again $e' e$, and this one is nothing but SST total square root of this.

So, this can be written as $e' e$ by SST total square root and $e' e$ is nothing but SSE residual by SST , and SSE residual is nothing but SST total minus SSR regression by SST . So, we are almost done, and then this can be written as $1 - SSR$ by SST , so this one is nothing but $1 - R^2$, so $1 - R^2$, so we proved that the correlation between e and Y is square root of $1 - R^2$.

(Refer Slide Time: 14:18)



So, here is the problem, so we have solved part one of this problem and what is the implication of this, the implication of this result is that, it is a mistake to attempt to find the defective regressions by a plot of residual e_i against the observation Y_i as this always shows a slope. So, if you can recall, you know once we have a fitted model say \hat{Y} is equal to $X\beta$, then what we do is that we compute the residuals and in a topic called model adequacy checking, we talked about several residual plots.

And the residual plots is sort of to check, whether the model assumptions are correct or also to check the goodness of the fit. So, what we do in the residual plot is that we plot, e_i residual against \hat{Y}_i not Y_i , so this is the reason why we plot the residual against \hat{Y}_i not against Y_i , because there is a correlation between e_i , and there is a correlation between e and Y . And the correlation, we just proved that it is $1 - R^2$.

root of that, so that is why we do not go for plotting e_i against Y_i , because there is always a theoretical slope between them. Since, because of this correlation, now so it the second part says that this slope is 1 minus R square, what is the meaning of this one, is that if you fit linear relationship between e and Y , the slope is going to be 1 minus R square in that linear fit.

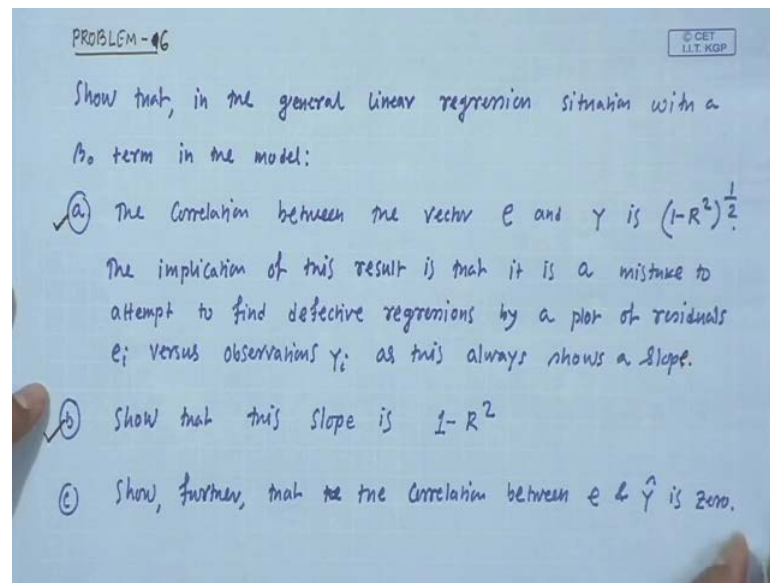
(Refer Slide Time: 17:11)

The image shows a handwritten derivation on a blue background. It starts with the equation $e, Y \mid e = a + bY$. To the right, the regression equation is given as $y = \beta_0 + \beta_1 X$ with the slope $\beta_1 = \frac{S_{xy}}{S_{xx}}$. The derivation then shows the slope b as $b = \frac{e'Y}{Y'Y} = \frac{e'e}{Y'Y}$. This is further simplified to $b = \frac{SS_{Res}}{SS_T} = \frac{SS_T - SS_{Reg}}{SS_T} = 1 - \frac{SS_{Reg}}{SS_T} = (1 - R^2)$. A hand holding a pen is visible at the bottom right of the image.

So, now let me fit relation between e and Y , suppose the relation is it is a straight line relation, so e is equal to say a plus bY . So, what I have to do is that I have to prove that this b is equal to 1 minus R square, the slope is 1 minus R square, this is part b. So, we know what is this b is equal to e prime Y by Y prime Y , because you know when we fit like Y equal to β_0 plus $\beta_1 X$, we know that β_1 is nothing but S_{XY} by S_{XX} .

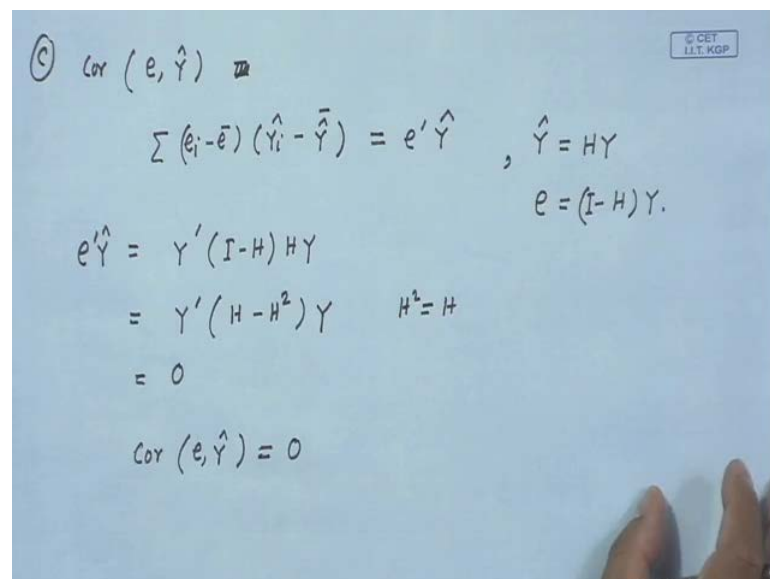
So, similar to that, so S_{XY} is nothing but X prime Y and in X prime X , so this has been followed from here, now e prime Y is just now we proved that e prime Y can be written as e prime e , so by Y prime Y . So, e prime e is nothing but $SS_{Residual}$ by Y prime Y is SS_{Total} , so this one is nothing but SS_{Total} minus $SS_{Regression}$ by SS_{Total} , so this is nothing but 1 minus $SS_{Regression}$ by SS_{Total} , which is equal to 1 minus R , so the slope the b we have proved that this is 1 minus R square.

(Refer Slide Time: 19:16)



And the third part of this question is that, so further that the correlation between e and \hat{Y} is equal to 0 say in the residual plot, we plot e against Y i \hat{Y} , because their correlation is 0. We have to prove that, now may be I proved this in the during the class also, anyway so the correlation between part c between e and \hat{Y} and that one is equal to.

(Refer Slide Time: 19:47)



So, first we will start with summation e_i minus \bar{e} into \hat{Y}_i minus $\bar{\hat{Y}}$, now \bar{e} is equal to 0. So, this can be written as e' \hat{Y} , and we know that \hat{Y} is

equal to $H Y$ and also we know that e is equal to 1 minus H , I minus $H Y$, just now we proved this two things.

Now, if I plug this two values here, what I will get is that my e prime Y hat is equal to e prime, so Y prime I minus H and Y hat that is $H Y$, so this is equal to Y prime H minus H square into Y , now see this H is a Hadamard matrix. So, H square idempotent matrix, so H square is equal to H , that is why this is equal to 0 , so the correlation you can see the covariance is equal to 0 . So, the correlation is the numerator this is the numerator of this correlation expression, so the correlation between e Y hat is equal to 0 , because of this fact, so we are done with the first problem.

(Refer Slide Time: 22:59)

PROBLEM-7

PROVE that the multiple correlation coefficient R^2 is equal to the square of the correlation between Y & \hat{Y} .

$$[Cor(Y, \hat{Y})]^2 = R^2 = \frac{SS_{reg}}{SS_T}$$

$$r_{Y\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{\hat{Y}})^2}}$$

$$= \frac{\sum (\hat{Y}_i - \bar{\hat{Y}})(\hat{Y}_i - \bar{\hat{Y}}) + \sum (\hat{Y}_i - \bar{\hat{Y}})e_i}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{\hat{Y}})^2}}$$

$Y_i = \hat{Y}_i + e_i$
 $\sum e_i = 0$
 $\sum (Y_i - \bar{Y}) = \sum (\hat{Y}_i - \bar{\hat{Y}})$
 $\bar{Y} = \bar{\hat{Y}}$

Let me prove one more problem, which is again you know theoretical call it problem 7. This problem says that you have to prove the coefficient of determination is equal to the square of the correlation between Y and Y hat. So, what you have to prove here is that, you have to prove that the correlation between Y and Y hat and the square of this one is equal to R square, R square is again the coefficient of determination, which is equal to S regression by S total.

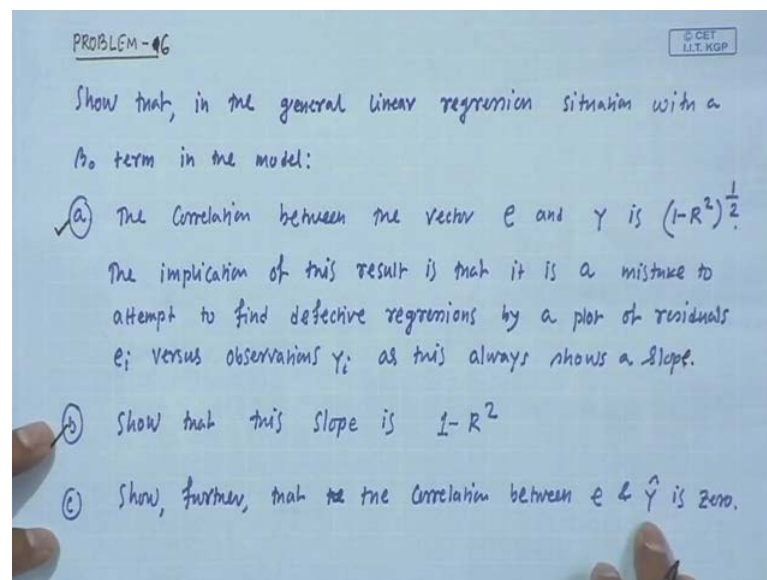
So, it is nice to prove this one, let me write down that what is R this correlation, I will use the notation $R_{Y \hat{Y}}$, this is the correlation between Y and Y hat which is equal to summation Y_i minus Y bar into Y_i hat minus Y hat bar. You can write, but this two are same by summation Y_i minus Y bar square summation Y_i hat minus Y bar, I hope you

understand that \bar{Y} is equal to $\bar{\hat{Y}}$ this is the mean of the observed value and this is the mean of the estimated value.

You know that $\sum e_i$ is equal to 0, for a model with intercept, and then e_i is nothing but $Y_i - \hat{Y}_i$, so this says that $\sum Y_i - \sum \hat{Y}_i$ is equal to 0, and then of course \bar{Y} is equal to $\bar{\hat{Y}}$ anyway, so this square the square root of this. Now, this can be written as we know that $Y_i = \hat{Y}_i + e_i$.

So, if I put that expression here what I will get is that, I will get $\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2$, because this is nothing but $\sum \hat{Y}_i^2 - 2\bar{Y}\sum \hat{Y}_i + \sum \bar{Y}^2$, because of this plus $\sum e_i^2$, so plus $\sum \hat{Y}_i^2 - 2\bar{Y}\sum \hat{Y}_i + \sum \bar{Y}^2 + \sum e_i^2$. And the denominator is the same, let me write down that this is $\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2$. Now, what about this one is this 0, yes just now we have proved that the correlation between e and \hat{Y} is equal to 0, so this is the covariance between e and \hat{Y} , e and \hat{Y} .

(Refer Slide Time: 28:08)



Just, now we had proved that in the previous problem, we have proved that the correlation between e and \hat{Y} is 0; that means the covariance between e and \hat{Y} is equal to 0.

(Refer Slide Time: 28:19)

PROBLEM-7

PROVE that the multiple correlation coefficient R^2 is equal to the square of the correlation between Y & \hat{Y} .

$$[Cor(Y, \hat{Y})]^2 = R^2 = \frac{SS_{Reg}}{SS_T}$$

$$r_{Y\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{\hat{Y}})^2}}$$

$$= \frac{\sum (\hat{Y}_i - \bar{\hat{Y}})(\hat{Y}_i - \bar{\hat{Y}}) + \sum (\hat{Y}_i - \bar{\hat{Y}})e_i}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{\hat{Y}})^2}}$$

$Y_i = \hat{Y}_i + e_i$
 $\bar{Y} = \bar{\hat{Y}}$
 $\sum e_i = 0$
 $\sum (Y_i - \hat{Y}_i) = 0$
 $\sum Y_i = \sum \hat{Y}_i$
 $\bar{Y} = \bar{\hat{Y}}$

So, this term is going to be equal to 0, so what we are left with is that then the correlation this $R_{Y\hat{Y}}$ is equal to $\frac{\sum (Y_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$ by this.

(Refer Slide Time: 28:26)

$$r_{Y\hat{Y}} = \frac{\sum (\hat{Y}_i - \bar{\hat{Y}})^2}{\sum (Y_i - \bar{Y})^2} = \frac{SS_{Reg}}{SS_T} = \sqrt{R^2}$$

$$\boxed{(r_{Y\hat{Y}})^2 = R^2}$$

So, I can write this one as summation $Y_i - \bar{Y}$ square, and the square root of the whole thing is it clear, because this one is square, and then you could cancel out. So, $R_{Y\hat{Y}}$ is equal to this one, and this one is equal to the numerator is $SS_{regression}$ and the denominator is SS_{total} , so what I got is that this is equal to R^2 . So, that is what

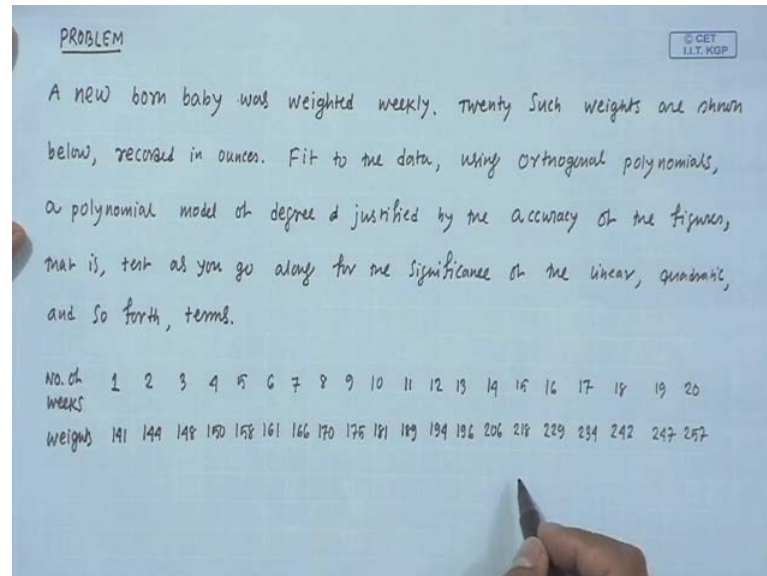
you have to prove that the square of the correlation between Y and \hat{Y} , so what you got is that $R_{Y\hat{Y}}^2$ is equal to R^2 , this is what we wanted to prove.

(Refer Slide Time: 29:57)

PROBLEM

A new born baby was weighted weekly. Twenty such weights are shown below, recorded in ounces. Fit to the data, using orthogonal polynomials, a polynomial model of degree d justified by the accuracy of the figures, that is, test as you go along for the significance of the linear, quadratic, and so forth, terms.

No. of weeks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Weights	141	144	148	150	158	161	166	170	175	181	189	194	196	206	218	229	234	242	247	257



So, next we will consider a practical problem, and this problem is involving orthogonal polynomial, I mean polynomial fitting using orthogonal polynomial. So, here in this problem we will sort of explain you know, how to decide about the degree of decide about the order of the polynomial. So, here the problem is that new born baby was weighted weekly, and 20 such weights are shown below, so this is the first week the baby has weight 141 ounce, and similarly for 20 weeks the data are given.

And if you plot a if you draw scatter plot for this one, I am sure that you are you are going to have a non-linear pattern may be. So, you fit to the data using orthogonal polynomials, a polynomial model of degree justified by the accuracy of the figures, so the degree is not given, so you have to decide about the degree of the polynomial, you are going to fit here.

So, we will sort of follow the that we talked about, how to decide about the degree while talking about orthogonal polynomial fit or polynomial fitting. So, you start with linear model, and then next you fit polynomial of order 2, and then you see the significance of beta 2 that is the coefficient of X^2 . If beta 2 is significant then only you go for third order polynomial, but if you see the beta 2 is not significant, then first order polynomial

first order model is enough, but if beta 2 is significant then you go for third order polynomial.

Again, you have to test the significance of beta 3, if beta 3 is significant, then you go for fourth degree polynomial, if it is not significant you stop at second degree polynomial something like that. So, here this problem sort of you know give idea about how to how to decide about the order of the polynomial, so let me recall little bit what is polynomial model.

(Refer Slide Time: 33:15)

we wish to fit the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \epsilon$$

$$\Downarrow$$

$$y = \alpha_0 + \alpha_1 P_1(x) + \alpha_2 P_2(x) + \dots + \alpha_k P_k(x) + \epsilon$$

$$SS_{Reg}(\alpha_1) = \hat{\alpha}_1 \sum y_i P_1(x_i)$$

$$SS_{Reg}(\alpha_2) = \hat{\alpha}_2 \sum_{i=1}^n y_i P_2(x_i)$$

$$SS_{Reg}(\alpha_3) = \hat{\alpha}_3 \sum_{i=1}^n y_i P_3(x_i)$$

$$\hat{\alpha}_0 = \bar{y}$$

$$\hat{\alpha}_j = \frac{\sum P_j(x_i) y_i}{\sum P_j^2(x_i)}$$

(x_i, y_i)

© CET IIT KGP

So, we wish to fit the model say Y equal to β_0 plus $\beta_1 x$ plus $\beta_2 x^2$ plus $\beta_k x^k$ plus ϵ . And what we learned there is that you know instead of fitting this model there are several advantage, if we go for polynomial fitting using orthogonal polynomial. So, instead of fitting this model, we fit Y equal to α_0 plus $\alpha_1 P_1(x)$ plus $\alpha_2 P_2(x)$ plus $\alpha_k P_k(x)$ plus ϵ , so $P_k(x)$ is k th order orthogonal polynomial and of course, they are all orthogonal polynomials.

So, we know how to fit this model we know what are these orthogonal polynomial, if you know I please refer my lecture on polynomial fittings, for to know more about you know this orthogonal polynomials. So, what we know is that, we know that $\hat{\alpha}_0$ is equal to \bar{y} , we know that $\hat{\alpha}_j$ is equal to $\frac{\sum P_j(x_i) y_i}{\sum P_j^2(x_i)}$. So, what I want to say here is that, so by using all this formula, so you know Y_i and of course, x_i are all you know equally spaced, you can just consider them 1 2 3 4 up to 20.

And you know this orthogonal polynomial, so this is the j th order orthogonal polynomial, so you can compute β_0 and β_j for j equal to 1 to k . Now, let me just write down that also you know that the SS regression due to α_1 , that is the contribution of the first order term in the polynomial, that is equal to $\hat{\alpha}_1 \sum_{i=1}^n Y_i p_{1x}$.

And similarly, this is regression due to α_2 is equal to $\hat{\alpha}_2 \sum_{i=1}^n Y_i p_{2x}$ for i equal to 1 to n , and similarly SS regression α_3 equal to $\hat{\alpha}_3 \sum_{i=1}^n Y_i p_{3x}$. Why I am talking about all this SS regression due to α_1 , α_2 , α_3 is that what you know I feel is that you first given the given the problem, you know you have x_i Y_i values.

(Refer Slide Time: 37:27)

PROBLEM

A new born baby was weighted weekly. Twenty such weights are shown below, recorded in ounces. Fit to the data, using orthogonal polynomials, a polynomial model of degree d justified by the accuracy of the figures, that is, test as you go along for the significance of the linear, quadratic, and so forth, terms.

No. of weeks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y_i : weights	141	144	148	150	152	161	166	170	175	181	189	194	196	206	218	229	234	242	247	257

$$SS_T = \sum (Y_i - \bar{Y})^2 = 26,018$$

So, these are the x_i Y_i values, this is your x_i , this is Y_i , so given this thing you first compute what is the total variability in the response variable. So, the response variable here is Y , so Y and this is x you compute SS total. First to check the variability in Y about the mean, that is $Y_i - \bar{Y}$ square, and once you know the total variability here it is you can check, that this is 26018..

So, this is the total variability and I want my model to explain this variability, now what I will do is that I will check the SS regression due to α_1 . So, SS regression due to α_1 will give me, how much of the total variability in Y is explained by the linear term. And similarly, the SS regression due to α_2 will give me, how much of the total variability is explained by the quadratic term that is by β_2 or α_2 , and

similarly S S regression due to alpha 3 will give the amount of variability, that is explained by the cubic term.

So, you see how much of the variability is explained alpha 1, how much of the variability is explained by the alpha 2, and how much of the variability is explained by alpha 3. If you see that alpha 1 and alpha 2 together they have almost explained major part of the variability in Y, then you can stop at the quadratic fit, but if you see that still there are significant part, which has not been explained by the quadratic fit, then you can go for the cubic term.

(Refer Slide Time: 40:12)

We wish to fit the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \epsilon$$

$$\Downarrow$$

$$y = \alpha_0 + \alpha_1 P_1(x) + \alpha_2 P_2(x) + \dots + \alpha_k P_k(x) + \epsilon$$

$$SS_{Reg}(\alpha_1) = \hat{\alpha}_1 \sum y_i P_1(x_i) = 25,438.75 \quad \left| \begin{array}{l} \hat{\alpha}_0 = \bar{y} \\ \hat{\alpha}_j = \frac{\sum P_j(x_i) y_i}{\sum P_j^2(x_i)} \end{array} \right.$$

$$SS_{Reg}(\alpha_2) = \hat{\alpha}_2 \sum_{i=1}^n y_i P_2(x_i) = 489.00$$

$$SS_{Reg}(\alpha_3) = \hat{\alpha}_3 \sum_{i=1}^n y_i P_3(x_i) = 1.15 \quad (x_i, y_i)$$

$$SS_T = \sum (y_i - \bar{y})^2 = 26,018$$

So, this is my S S total and I want to my model to explain this variability, so what I will do is that I will compute see I know, I am not going into detail of computing all this things. So, you can compute alpha 1 hat, you can check that this is equal to you can have the computations each you can check that this S S regression is equal to due to alpha 1 is very large. This is 25438.75 and S S regression due to alpha 2 is 489 0 0 and S S regression due to alpha 3 is 1.15, and let me repeat that S S total that is Y i minus Y bar the variability in the response variable about mean is equal to 26018.

So, you can see this is the total variability, and the first order term or the linear term has already been explained major part of this total variability, and alpha 2 is also significant this is 489, and what you can do is that you know see you can now make a anova table, let me do that.

(Refer Slide Time: 41:48)

ANOVA TABLE				
Source	df	SS	MS	F
Reg (α_1)	1	25,438.75	25,438.75	4558.92
Reg (α_2)	1	489.00	489.00	87.63
Reg (α_3)	1	1.15	1.15	0.21 < 4.49
Res	16	89.30	5.58	
<hr/>				
Total	19	26,018		

$$Y = 136.227 + 2.68X + 0.167X^2 \quad \left| \begin{array}{l} F \\ .05, 1, 16 \end{array} \right. = 4.49$$

So, anova table source degree of freedom S S M S and F, so regression due to alpha 1, regression due to alpha 2, regression due to alpha 3 and say residual and then total. So, total S S we computed that this is 26018, and S S regression is due to alpha 1 is 25438.75, S S regression due to alpha 2 is 489, and S S regression due to alpha 3 is 1.15. And you can check that the S S residual is just 89.30 and the degree of freedom for this one is 1, this one is 1, this one is 1, and the total degree of freedom is 19, because there are 20 observations, and then the residual degree of freedom is 16.

So, you can compute the M S value, M S residual is 5.58, and this value will be will remain same 25438.75 489.00 1.15. Now, see of course, then the a value F value, you can check that this value by this, that is going to be 4558.92 and F value for alpha 2 is this by this that is 87.63, and the F value for alpha 3 is 0.21, and this all this F follows has degree of freedom 1 16 and now you check the tabulated value of 0.05 1 16 that is equal to 4.49.

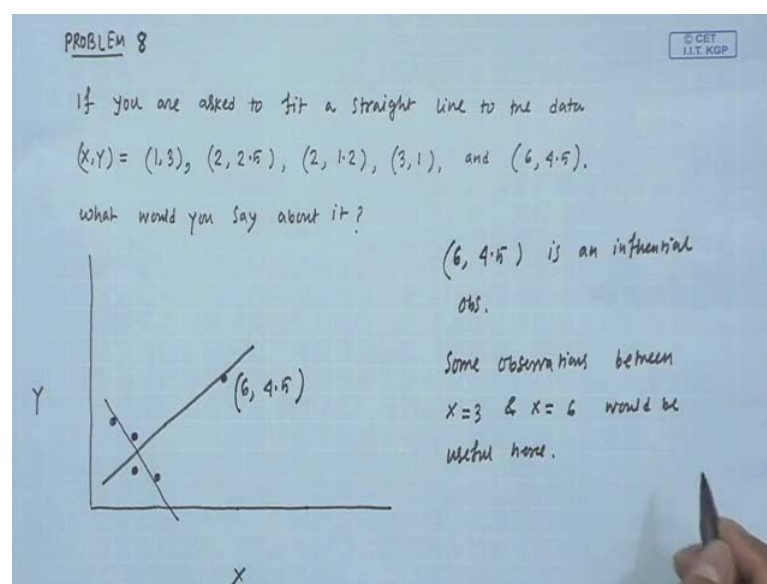
So, you can see that this one is much larger than the tabulated value, this one is also much larger than the tabulated value, but this one is smaller than 4.49. The meaning of this one is that alpha 1 is significant, alpha 2 is also significant, because the f value observed F value is greater than the tabulated F value, and this table says that alpha 3 is not significant; that means, you can go for a quadratic fit.

So, the final fit in terms of x is 136.227 plus $2.68x$ plus $0.167x^2$, so that you can check, but what I want to say is that this is my total variability, and I want to explain I want my model to explain these variability. Now, you can see that this linear term explain huge part of the total variability, the quadratic term explain this much of total variability, and the cubic term is not significant.

And now, we can see that you know out of $26,000$ of the total variability major part has been explained by the model and the remaining part is 89.30 . So, I can go for a quadratic fit, but here just I want to say, suppose your cubic term is not significant, but this residual is still reasonably high, then you can go for α^4 see α^3 might not be significant, but α^4 could be significant I mean I feel, so.

So, in that case you can go for a with linear term, quadratic term and the fourth degree, but not the third degree, so it depends on how small the residual is if the residual is reasonably small. Then you do not need to go for the higher order polynomial, but somehow I feel that you know, even if the third degree, even if the cubic term is not significant, it might happen that the fourth degree term is significant. But, you have to go for the fourth degree term fourth degree polynomial, only if the residual part is large otherwise, you know just forget about that you go for the quadratic fit. So, this is the problem, we talked about from the polynomial regression, and next let me talk about one more problem.

(Refer Slide Time: 48:12)



This is I do not know, what is the number this might be 8, so this problem says that, if you are asked to fit a straight line to the data, this given data what would you say about it. I mean it is a sort of difficult exactly know what the question is asking for or looking for. So, you are given 1 2 3 4 5 data points and here is the scattered plot, for the given data, so this is my x axis, this is my Y axis and here do we have anything to say, because see here you can see that this particular point that is 6 4.5.

This particular point is not in the usual trend of the data, so that means at some point of time you know, I talked about influential observation and leverage point. So, this point you know this 6 4.5 it seems to be a influential point, because this one is not in the general trend of the other data, so 6 4.5 is an influential observation. And see, if you with this point I mean without this point you can you can think of fitting, if you fit a model to the remaining data you will get a fit with negative slope, so this should be the fitted mode line, I mean you will fit a line like this, for if you ignore this point.

And, but if you include this point then your fitted line should be something like this one, so that is the problem with influential observation. So, without the point this 6, this is my 6 4.5, so without this point the fitted model has negative slope, and with this point the fitted model has positive slope. So, this clearly says that the point 6 4.5 is an influential observation, so the recommendation here is that you know, if there exist influential observation very and few influential observation.

If you can identify an influential observation in the given data may be you can ignore them, if they are not large in number you can ignore them, and fit a model for the remaining data. And what could have been better here is that you know, if you see here you know you have data up to 3, and then the next one is 6. So, some observations between x equal to 3, and x equal to 6 would be useful here, so this is also you know you can comment in that way. So, we solved you know some problems in this tutorial, and again in the next tutorial we will solve some randomly selected problems, now we need to stop.

Thank you.